Towards High-Fidelity Face Normal Estimation

Meng Wang*
autohdr@gmail.com
Tianjin University

Chaoyue Wang
wangchaoyue9@jd.com
JD Explore Academy

Xiaojie Guo
xj.max.guo@gmail.com
Tianjin University

Jiawan Zhang†
jwzhang@tju.edu.cn
Tianjin University

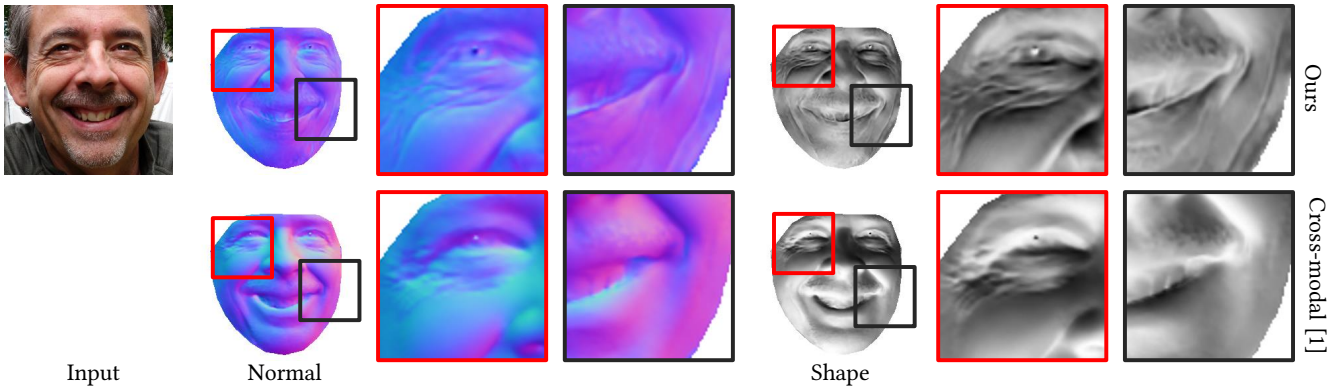

Figure 1: With normal features injecting into face structure features in our network, our model enables, for the first time, to produce high-fidelity face normal by example-based learning.

ABSTRACT

While existing face normal estimation methods have produced promising results on small datasets, they often suffer from severe performance degradation on diverse in-the-wild face images, especially for the high-fidelity face normal estimation. Training a high-fidelity face normal estimation model with generalization capability requires a large amount of training data with face normal ground truth. Since collecting such high-fidelity database is difficult in practice, which prevents current methods from recovering face normal with fine-grained geometric details. To mitigate this issue, we propose a coarse-to-fine framework to estimate face normal from an in-the-wild image with only a coarse exemplar reference. Specifically, we first train a model using limited training data to exploit the coarse normal of a real face image. Then, we leverage the estimated coarse normal as an exemplar and devise an exemplar-based normal estimation network to explore robust mapping from the input face image to the fine-grained normal. In this manner, our method can largely alleviate the negative impact caused by lacking training data, and focus on exploring the high-fidelity normal contained in natural images. Extensive experiments and ablation studies are conducted to demonstrate the efficacy of our design, and reveal its superiority over state-of-the-art methods in terms of both training data requirement and recovery quality of fine-grained face normal. Our code is available at https://github.com/AutoHDR/HFFNE.

CCS CONCEPTS

• **Computing methodologies** → **Image manipulation**.

KEYWORDS

Face normal estimation, High-fidelity, Exemplar-based learning

ACM Reference Format:
Meng Wang*, Chaoyue Wang, Xiaojie Guo, and Jiawan Zhang†. 2022. Towards High-Fidelity Face Normal Estimation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547959

1 INTRODUCTION

Surface normal estimation for facial images is an important intermediate component in understanding 3D face structure in multimedia, computer vision and graphics, which has a wide range of applications, such as augmented reality, virtual reality and 3D face modeling. The problem is highly ill-posed, as infinite recoveries from an image are feasible and it is difficult to determine which one is correct without extra constraints.

Over last decades, a variety of approaches have been proposed to tackle face normal estimation by learning to recover facial components from a single image. Most of these methods assume that key geometric information about human faces is contained in the 2D image. With the prior knowledge learned from training dataset, a well-trained model is capable of recovering face information (e.g., face normal) from an input image [40, 41, 47, 54]. Recently, Abrevaya *et al.*[1] made a further step, which treats the face normal estimation as a cross-domain/cross-modal image translation problem. Benefited from the powerful fitting ability of deep neural

* This work was performed when Meng Wang was visiting JD Explore Academy as a research intern.
† Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547959

networks on paired training data, the proposed 'Cross-modal' [1] works well for face geometric structure and achieves convincing performance on test data.

Although satisfactory results under certain situations, the following challenges remain: 1) Uneven distribution between synthetic data and real data; 2) Lack of a large amount of training data; and 3) loss of high-frequency geometric details. According to our observation, directly learning the high-fidelity translation mapping from image pixels to face normals largely depends on the amount and quality of training data, and the well-trained model is vulnerable to different environments, such as fluctuations of ethnicities or poses. During training, the lack of consideration of multi-scale face structure features makes the estimated face normal short of high-fidelity geometric details.

The thought of improving the cross-domain mapping robustness motivates us to design a new model for mitigating the preparation of training data and better exploit the domain relationship for the more accurate normal estimation. Towards this purpose, we design a simple yet efficient solution to overcome the aforementioned limitations without requiring large-scale training data neither generated from scanned 3D data nor from massive public ground truth data. Specifically, we propose a framework to leverage the effectiveness of exemplar-based deep learning on the task of normal estimation from a single in-the-wild image. In our framework, we first train an estimation network on a small dataset to produce coarse normal from input facial images. Having the estimated coarse result as an exemplar, we further customize an exemplar-based mapping network to produce high-fidelity normal recovery from the input face image. Different from previous cross-domain translation methods, our network is logically divided into three sub-networks, including an exemplar encoding network, a face encoding network, and a feature injection network. The exemplar encoding network encodes the exemplar to an intermediate latent representation where the reliable normal features can be established. The face encoding network learns the face geometric structure features, while the feature injection network employs a set of feature modulation modules from StyleGAN2 [29] to synthesize the high-fidelity normal by modulating feature weights. The modulation focuses on global and local face geometric information for successive generations of the final high-fidelity normal. These three sub-networks facilitate each other and are learned with simple perceptual and reconstruction losses in an end-to-end manner. Our method produces high-fidelity normal (see Fig. 1) and outperforms previous methods in normal quality by a large margin, with an unfaithful normal as reference. The major contributions of this work can be summarized as follows:

- Inspired by the exemplar-based learning, our proposed network is able to train with coarse exemplars when reliable references are unavailable, instead of using any ground truth.
- We design a framework that estimates high-fidelity face normal from a single in-the-wild image by optimizing face and normal feature injection, which is of strong generalization/transfer abilities, even though the network is trained purely on a inconsistently distributed dataset.
- Extensive experiments together with ablation studies are conducted to demonstrate the efficacy of our design and its superiority over state-of-the-art alternatives.

2 RELATED WORK

2.1 Surface normal estimation

Shape from shading (SfS) [17] is a popular strategy for image-based 3D surface reconstruction based on shading cues, which plays an important role in recovering geometry. Traditionally, Shape from shading is always considered as an optimization problem under a Lambertian shading model [4–6, 49, 51]. For example, Barron *et al.*[6] utilize a series of priors respectively on shape, reflectance, and illumination, and design a multiscale optimization technique to seek the shape. Xiong *et al.*[49] propose a framework based on a quadratic representation of local shape to recover accurate local shape and lighting. Ecker *et al.*[13] design a polynomial system to solve SfS problem for polyhedral and curved surfaces without requiring boundary conditions. In order to maintain analytical tractability, all these methods make substantial assumptions that may not always hold in unconstrained settings. For instance, Barron *et al.*[6] assume a known object boundary, and Xiong *et al.*[49] assume quadratically parameterized surfaces, both of which are typically unavailable in practice.

Recently, data-driven methods [40, 41] combined with SfS have been studied for normal estimation. Shu *et al.*[41] build an end-to-end generative network that infers a face-specific disentangled representation of intrinsic face components, like normal (shape). Sengupta *et al.*[40] propose a two-stage training strategy to learn low-frequency variations from synthetic data in the first stage, and in the second stage, synthetic labeled data and unlabeled real-world images are trained together to learn high-frequency details from real images through the photometric reconstruction loss. However, the smooth constraints in [41] make the results lacking high-frequency details. And the prior/knowledge solely learned from synthetic data in [40] likely hinders these methods from practical applications, due to the gap between the synthetic and real data.

Closely related to our work are methods that recover face normals from an image using deep neural networks, *e.g.* [9, 14, 24, 45, 46, 48, 63]. All of these methods estimate the face normal when recovering the 3D information, rather than estimating normal in a targeted manner. Although they are able to recover the normal, a large room for improving the quality of high-frequency details exists. To address this problem, Tran *et al.*[44] utilize a dual-pathway network to learn additional proxies as means to side-step strong regularizations. Their approach focuses on model recovery, and the pertained model relies on a synthetic facial mesh. Alternatively, our work enables exemplar-based learning and only needs coarse normal as exemplar, which significantly broadens the applicability to scenarios with limited training data available.

2.2 Cross-domain tasks

Many computer vision problems, such as style transfer [15, 33], image inpainting [53, 55], and image colorization [8, 11] to name just a few, can be considered as cross-domain (cross-modal) learning tasks [10, 52]. The cross-domain learning is essentially to map an input image from one domain to a target one. Exemplar-based learning is a kind of cross-domain method, which uses a content image and an exemplar image to generate the target image contained both content from the input and style from the exemplar.

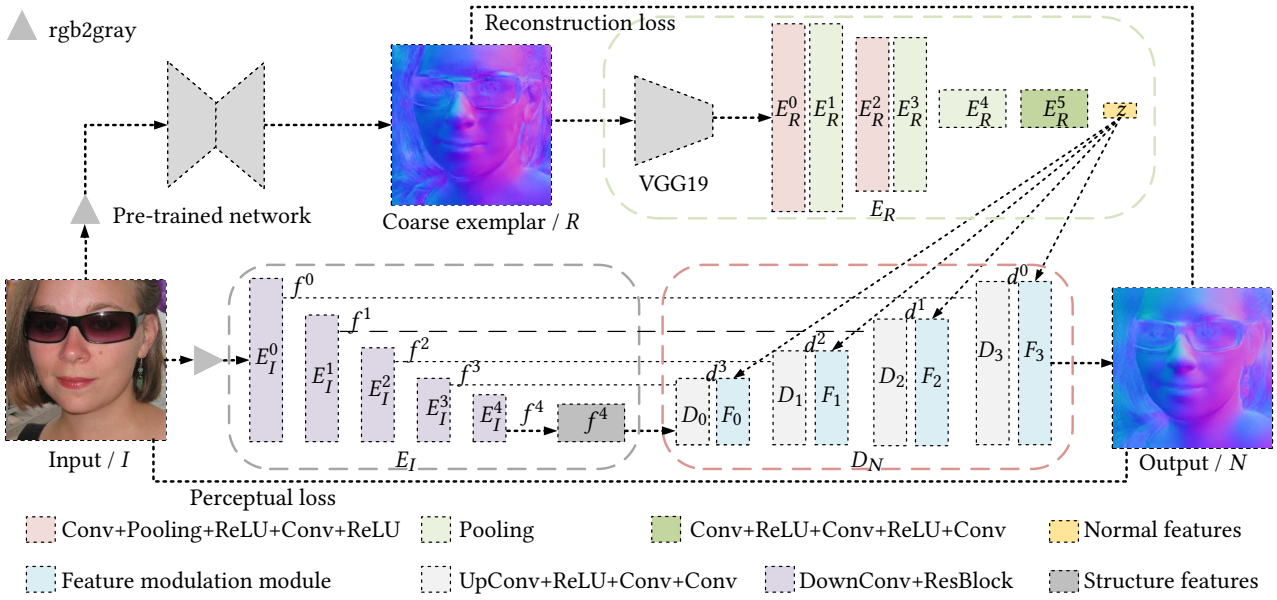

Figure 2: Illustration of our coarse-to-fine framework, which can produce high-fidelity normal via encoding the coarse exemplar normal into normal features together with the face structure features.

It has recently achieved steady progress with the help of convolutional neural networks and has been widely used for various image synthesis tasks, such as image colorization [16, 50, 58], image translation [27, 57, 59], image super-resolution [12, 18, 62], and image inpainting [22, 36].

In the literature, few research efforts focused on exemplar-based surface normal estimation can be found. It is worth mentioning that Huang *et al.* [20] propose an exemplar-based approach that estimates surface normals from a single image. However, their database is synthesized from known 3D models and the optimization is to find the most likely normals in the database. With the emergence of deep neural networks, elaborately designed networks can better extract image features to represent high-frequency details. Abrevaya *et al.* [1] propose a cross-modal method for face normal synthesis and design an architecture that enables face details to be transferred between the image and normal domains with deactivable skip connections. But, both of the mentioned methods loss the high frequency details on the estimated normal and need synthesized data. Our method largely alleviates the requirement of data acquisition. During inference, we only need the input facial image and its coarse exemplar normal as inputs.

3 METHOD

Our goal is to predict the high-fidelity face normal giving a still facial image. Specifically, a pre-estimated coarse normal is applied as a not aligned normal in face. Then, a certain mapping from the face structure features and normal features to the geometric space is learned. Finally, both of the coarse normal and the learned mapping fall back to a plausible high-fidelity face normal. However, without knowing the distribution of face normal, it is difficult to accurately estimate face normals with deep learning network. Regard this issue, we utilize the Photoface dataset [56] as ground truth to train a coarse normal estimation model that aims to capture the possible (inaccurate) distribution of face normal. Additionally, a model trained on limited data will hardly perform well on diverse in-the-wild face images, due to the domain gaps between the training data and various real-world facial images. In our model, the coarse normal estimated by the pre-trained coarse network is employed as reference (coarse exemplar). To alleviate the difference between coarse estimations and fine-grained outputs, we adopt the smooth-L1 loss [21] as the distance metric to relieve the ambiguity. Moverover, it is challenging to propagate coarse exemplar features properly to face structure features based on hand-crafted rules for the sake of generating fine-grained normal maps from the normal domain. To address this challenge, we introduce a feature modulation module [29] as our backbone to learn the representation of normal features in the face. Details of the whole framework are explained in the following subsections.

3.1 Coarse-to-fine prediction

As illustrated in Fig. 2, our framework first employs a pre-trained network to generate a coarse exemplar normal $R \in \mathcal{R}^{H\times W\times 3}$ from an input face image $I \in \mathcal{R}^{H\times W\times 1}$. The fine-grained normal $N \in \mathcal{R}^{H\times W\times 3}$ can be generated by the learned mapping function $\phi_{(I,R)\rightarrow N}$ from feature spaces of I and R. The high-fidelity face normal estimation N is conditional on both the face I and the coarse exemplar R, which can be formally represented as $N = \phi(I, R)$.

The whole framework contains three sub-networks, including a normal feature encoder E_R, a facial features encoder E_I, and a normal feature decoder D_N as depicted in Fig. 2. Specifically, E_R takes a coarse exemplar R^i as input and generates the coarse normal features $\mathbf{z}^i$ using three convolution blocks and three pooling layers.

The normal features encoded from the coarse exemplar can be expressed as:

$$\mathbf{z^i} = E_R(R^i). \tag{1}$$

In addition, E_I represents the high-fidelity face content encoder network that contains five downsampling convolutions to extract the high-fidelity face content features $\mathbf{f_l}$ at the convolutional block $\mathbf{l}$ as follows:

$$\mathbf{f_l} = E_I^{\mathbf{l-1}}(\mathbf{f_{l-1}}). \tag{2}$$

The intermediate features are passed to the fine-grained normal decoder D_N, and provide multi-scales structure information. The decoder D_N contains consecutive upsampling blocks and feature modulation modules F. The normal decoder D_N executes as:

$$\mathbf{d_i} = \begin{cases} F_i\left(D_i\left(\mathbf{f_N}, \mathbf{f_{N-1}}\right)\right), & \text{if } \mathbf{i} = N-1 \\ F_i\left(D_i\left(\mathbf{d_{i+1}}, \mathbf{f_i}\right)\right), & \text{otherwise.} \end{cases} \tag{3}$$

These high-fidelity structure features $\mathbf{f_l}$ are modulated with the normal features $\mathbf{z^i}$ to produce multi-scales normal features, which can be decoded by D_N to produce the fine-grained normal.

3.2 Feature modulation module

As mentioned in [26, 29, 61], the feature modulation module can decide the desired feature weights in a learnable way. To make full use of structure features and normal features, which respectively offer high-fidelity facial details and normal distribution, we introduce the feature modulation module [29] into our normal decoder network. The decoder implicitly modulates weights of normal features affected by content features with a multi-scales injection in the feature space. It can be formulated as:

$$\bar{\mathbf{w}} = \mathbf{w} \cdot \mathbf{s} \cdot F_{\text{Linear}}(\mathbf{z}), \tag{4}$$

where $\bar{\mathbf{w}}$ and $\mathbf{w}$ are the modulated convolution weights and original weights, respectively. Both of them are $\mathcal{R}^{C_i \times C_j \times K \times K}$, with K, C_i and C_j being the kernel size of the weights, the numbers of input channels and output channels, respectively. The F_{Linear} injects the normal distribution features $\mathbf{z^i}$ into weights in this features scale $\mathbf{s}$, which is the scale corresponding to the feature map. After the modulation, we adopt F_{Norm} normalization to further restrict the values of convolution weights, the normalization is formulated as:

$$F_{Norm} = \frac{\bar{\mathbf{w}}}{\sqrt{\sum \bar{\mathbf{w}}^2 + \epsilon}}, \tag{5}$$

where ϵ is a small positive constant for avoiding zero denominator. To restore the outputs back to unit standard deviation, we also normalize the dimension of $\bar{\mathbf{w}}$. The final convolution weights are determined as:

$$\hat{\mathbf{w}} = F_{Norm}\left(\bar{\mathbf{w}}\right). \tag{6}$$

Given facial structure features $\mathbf{f}$, the modulated features $\mathbf{m}$ can be written as,

$$\mathbf{m} = F_{conv}\left(\hat{\mathbf{w}}, \mathbf{f}\right), \tag{7}$$

where F_{conv} is a convolution operation. We have now baked the entire normal features to a single convolution layer.

3.3 Architecture

Skip-connection based encoder-decoder networks are common and simple for the task of image-to-image translation [23, 37]. We observe that most of the high-frequency variations are passed from the encoder to decoders through the skip connections. Thus, we consider two network architectures based on skip-connection that can produce promising results without requiring an elaborated design and demonstrate the effectiveness of our approach.

In this paper, we design three different architecture configurations for feature injection: (***A1***), the exemplar is used as a condition for the input face and is directly concatenated via a simple skip-connection between blocks in the encoder and decoder. (***A2***), the features of input face structure and exemplar are concatenated at the feature space for further prediction. The results shown in the first two rows of Table 3 suggest that a simple adjustment of the network can achieve better results. Furthermore, we explore the performance by adding a feature modulation module for the injection of structure features and normal features (***A3***). The feature modulation module is inserted at every layer of the decoder D_N. With the feature modulation module, the results are able to achieve competitive performance (see the last row of Table 3). Due to limited space, the detailed architectures of (***A1***) and (***A2***) can be found in the supplementary material.

3.4 Loss function

Our model is capable of producing high-fidelity face normal without need a large amount of ground truth. Moreover, the facial details should represent the accuracy of normal direction in the corresponding regions. To accomplish these objectives, we employ the following loss terms:

Reconstruction loss. The fine-grained output normal is desired to be similar to the coarse exemplar normal, yet they should not be exactly same in details/outliers. To make the model more robust, we adopt the smooth-L1 loss [21] between the fine-grained output and the coarse exemplar normal as follows:

$$\mathcal{L}_{\text{recon}} = \sum \text{smoothL1}\left(N_c, N_r\right), \tag{8}$$

where N_c and N_r designate the coarse exemplar normal and the estimation of fine-grained normal, respectively.

Perceptual loss. Solely adopting the smooth-L1 loss will result in output normals similar to even the same as the coarse exemplar normal. Thus, this will not serve the purpose of fine details. Here, we add a perceptual loss term to mitigate this problem. The k-th layer of an off-the-shelf image encoder e (VGG19 [42]) predicts a representation $e^{(k)}(\mathbf{I}) \in \mathbb{R}^{C_k \times W_k \times H_k}$ (in this work, k is 'relu1_2'). The perceptual loss is given by:

$$\mathcal{L}_{\text{perc}} = \sum_k \left\| e^k(I) - e^k(N_r) \right\|_1, \tag{9}$$

where $\|\cdot\|_1$ means the L1 norm.

In summary, the overall loss function for training is defined as :

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ref}} \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{perc}}, \tag{10}$$

where λ_{ref} is the weight of the normal reconstruction term. In the experiments, we empirically set $\lambda_{\text{ref}} = 1$, which works sufficiently well.

Figure 3: Coarse exemplar (co-exemplar) vs our high-fidelity normal (HF-normal) on several samples from the FFHQ dataset [28].

3.5 Discussion

Joint training of two steps. We encode a coarse exemplar normal as normal features to our fine-grained network. Why we train the framework in two steps? The simple way is to train with ground truth normal, but the model relies on the distribution of the training dataset, resulting in poor generalization ability in the wild, as shown in Fig. 3 and Fig. 8. Considering diverse face normals, various poses and expressions, our pipeline could also be separately trained in two steps. In the first step, a coarse normal is generated as a reference, and then the second step devotes to produce the fine-grained normal under the guidance of coarse normal. The division of Labour is clear-cut, each step being charged with specific responsibilities. By this way, we have a better generalization ability and a faster convergence speed on new samples

Feature injection with relatively low dimensions. We first encode a coarse exemplar normal into low-dimensional features which are supposed to represent a normal distribution for one face. The benefits of doing so are twofold. First, as shown in Fig. 2, during training, given a coarse exemplar, we extract the normal features, which are not aligned in face structure but contain the geometric information of the input face. Second, to produce the fine-grained normal, we do not hope the model learning to simply copy the coarse normal to its output (considering the reconstruction loss), such as (***A1***). In contrast, coarse normal features only provide the guidance (*i.e.* exemplar) of output normal, yet most of details are learned from the input image with high-fidelity details.

4 EXPERIMENTS

4.1 Setup

Datasets. We test our method on six face datasets, including the 300-W [38], CelebA [32], FFHQ [28], Photoface [56], Florence [2] and ICT-3DRFE [34] datasets. The 300-W dataset consists of 300 indoor and 300 outdoor in-the-wild images. The CelebA is a large-scale real face dataset in the wild. The FFHQ contains a wide range of ages and ethnicities. Each case of the Photoface contains a set of images with four different lightings, which can use a photometric stereo method to estimate normal as ground truth. Since the authors do not provide the training split of the dataset, following the setting in [1, 40, 47], we create a random split and collect about 20 percent of image/normal pairs (about 2.5k) for our evaluation. The rest image/normal pairs are used to train the model to produce our coarse exemplar. In addition, we also generate face normals from 53 3D-models of the Florence dataset by following the work of [1]. This allows our model to evaluate on a completely unseen dataset. Considering the ICT-3DRFE dataset contains face albedos, we can employ the estimated normal and albedo to relight faces, which can demonstrate the accuracy of our method. Since the ICT-3DRFE is not publicly available, we downloaded low-resolution images from their webpage and then enhanced the images via face super-resolution [7].

Metrics. Following previous methods [1, 40, 47], metrics used for this task are the mean angular error between the output and the ground truth normals, and the percentage of pixels within the facial region with an angular error of less than 20°, 25° and 30°. In addition, we also adopt the geometric shading and normal error map for qualitative comparisons.

Implementation details. Our framework was implemented in PyTorch [35] with a learning rate 10^{-4}. Adam optimizer [30] was used with default parameters. Following [1], we also cropped the face with a fixed size and resized to 256×256. The pre-trained network was based on the generative networks architecture and a discriminator in [23]. We trained the pre-trained model about 200K iterations and refinement model about 150k iterations until the model converges with a batch size of 8. E_R firstly used the pretrained VGG19 [42] to extract deep convolution features from the layer of *'relu5_2'*.

4.2 Comparison

We compare coarse exemplars with the refined high-fidelity normals, as shown in the Fig. 3. Although the pre-trained model converges on the training dataset, its generalization ability is poor due

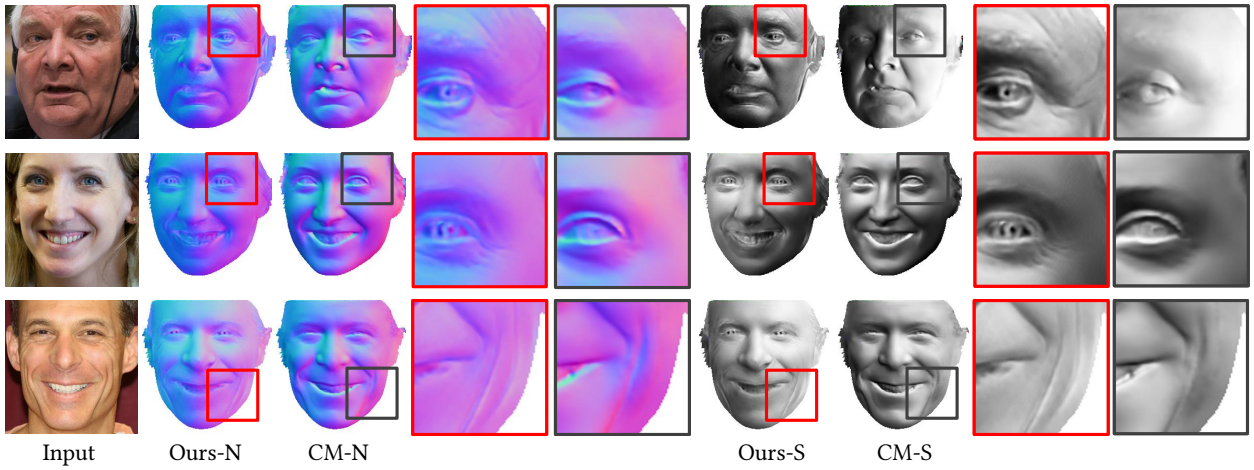

Figure 4: Comparisons in normal and geometric shading on the FFHQ [28]. We use different angles of light to render our geometric shading (Ours-S) and 'Cross-modal' [1] shading (CM-S) to exhibit more details.

Table 1: Normal reconstruction errors on the Photoface [56]. The lower mean error is better, while the higher is better for correct pixels at various thresholds.

Method	Mean ± std	< 20°	< 25°	< 30°
Pix2V [39]	33.9±5.6	24.8%	36.1%	47.6%
Extreme [43]	27.0±6.4	37.8%	51.9%	64.5%
3DMM	26.3±10.2	4.3%	56.1%	89.4%
3DDFA [63]	26.0±7.2	40.6%	54.6%	66.4%
SfSNet [40]	25.5±9.3	43.6%	57.5%	68.7%
PRN [14]	24.8±6.8	43.1%	57.4%	69.4%
Cross-modal [1]	22.8±6.5	49.0%	62.9%	74.1%
Ours-unpaired	**17.2±10.8**	**67.7%**	**79.8%**	**88.9%**
UberNet [31]	29.1±11.5	30.8%	36.5%	55.2%
NiW [47]	22.0±6.3	36.6%	59.8%	79.6%
Marr Rev [3]	28.3±10.1	31.8%	36.5%	44.4%
SfSNet-ft [40]	12.8±5.4	83.7%	90.8%	94.5%
Cross-modal-ft [1]	12.0±5.3	85.2%	92.0%	95.6%
LAP [60]	12.3±4.5	84.9%	92.4%	96.3%
Ours-paired	**11.3±7.7**	**88.6%**	**94.4%**	**97.2%**

to the distribution gap between the training dataset and the real data. The coarse exemplars produced by the pre-train model suffer from noticeable artifacts as can be seen in the second row. In contrast, our method is able to remove these artifacts and refine coarse exemplars to produce high-fidelity normals. To put it another way, our method effectively improves the generalization capability when facing limited training data.

In Table 1, we provide quantitative results by our reconstructed normals ('Ours-paired') in comparison with those by other state-of-the-art alternatives including 'Cross-modal-ft' [1], 'SfSNet-ft' [40], 'Marr Rev' [3], 'NiW' [47] and 'UberNet' [31]. All results given are trained and tested on input face images with resolution of 256×256. First, we compare ours with the methods that also trained on the Photoface [56]. Table 1 shows mean angular errors (degrees) and percentage of errors below $< 20°$, $< 25°$ and $< 30°$. '-ft' means that the method is fine-tuned on Photoface. 'Ours-unpaired' is trained using faces with unpair exemplar coarse normal while 'Ours-paired' is in a paired manner. Our methods improve normal estimation accuracy for all the degrees over the other methods.

Quantitative results on the Florence dataset [2] are shown in Table 2. Following the work of [1], we only compare the methods in Table 2 using the aligned output normal of face images for fair comparison. Our proposed model performs better in all metrics than the involved competitors. This validates that our method is more robust than others for out-of-distribution face images.

Table 2: Reconstruction error on the Florence dataset [2].

Method	Mean ± std	< 20°	< 25°	< 30°
Extreme [43]	19.2±2.2	64.7%	51.9%	64.5%
SfSNet [40]	18.7±3.2	63.1%	77.2%	86.7%
3DDFA [63]	14.3±2.3	79.7%	87.3%	91.8%
PRN [14]	14.1±2.2	79.9%	88.2%	92.9%
Cross-modal [1]	11.3±1.5	89.3%	94.6%	96.9%
Ours-Paired	**10.1±3.4**	**92.3%**	**95.6%**	**97.8%**

More importantly, given an input face image and an coarse exemplar normal, our proposed method produces a high-fidelity normal for the face, as shown in Fig. 5. 'Pix2V' [39] can also capture details about the face, such as wrinkles. However, these details are only available on large-scale angle changes and relatively poor on flat changes. 'SfSNet' [40] and 'Cross-modal' [1] are relatively smoothed in the recovery of normal and can accurately recover with large changes in angle. They are much better than 'Pix2V' [39] in terms of details and geometric variations in local places where the normal angle with small changes. However, they are failed to accurately estimate the normal where contains the fine-grained details of eyebrows, hairs and beards.

To compare in enhanced geometric shading, we show the normal and shading over the same base mesh obtained by PRN [14] in Fig. 4.

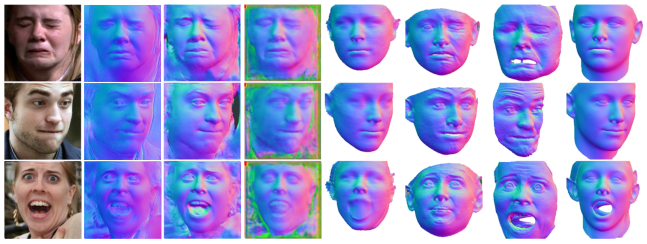

Figure 5: Normal comparison with the state-of-the-art methods on the data showcased by the 'Cross-modal' (CM) [1] on the 300-W dataset [38]. (Please zoom in for details, such as wrinkle, moustache or eyebrow.)

Figure 6: Normal error comparisons on the Photoface dataset [56]. 'GT', 'Ours-N', 'CM-N', 'Ours-E' and 'CM-E' are ground truths, our predictions, 'Cross-modal' [1] predictions, our error maps and 'Cross-modal' error maps, respectively.

Rendering under different angular lighting conditions makes it easier to observe the detailed information of the geometry. Compared to 'Cross-modal' [1], our normals recover much more fine-grained face geometric details that significantly enhance the base mesh. By using a coarse exemplar normal as reference with a perceptual loss, our network is able to generate high-fidelity normal that extends beyond the coarse exemplar subspace, better fits the shape of the input face, and exhibits more identity information.

In Fig. 6, we also show qualitative comparisons between ours and 'Cross-modal' [1]. It exhibits the normal estimations and normal error maps on test samples from the Photoface dataset [56]. The smaller the normal estimation error, the closer it is to 0 degrees and the darker the color of the error map. By using a coarse exemplar features injected into face structure features, ours can produce a more robust face normal estimations compared with 'Cross-modal'.

In Fig. 7, we show an application of normal estimation against 'Cross-modal' [1]. We first estimate face normal with albedo from the ICT-3DRFE [34]. Then we take a target light as input and generate a new relighting face with Lambertian reflectance [25]. We use ground truth normal as reference to show that our shading can ensure a more realistic shadow effect. Please refer to our supplementary for more comparisons.

5 ABLATION STUDIES

Initialization of coarse exemplar normal. We testify the influence of different initializations on coarse exemplar normals, as

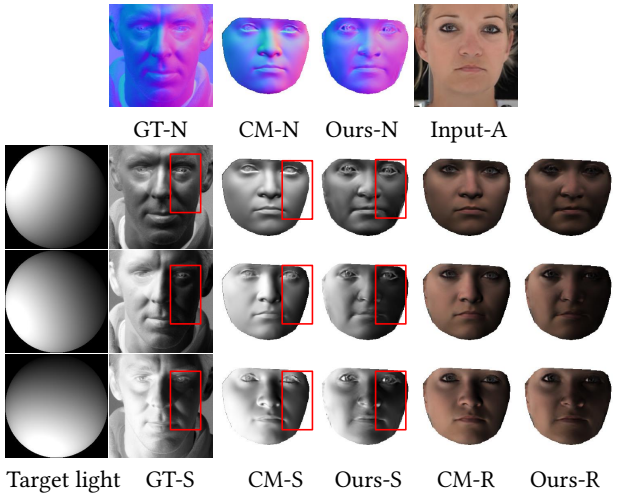

Figure 7: Face relighting with estimated normal. '-S', '-N', '-R' and '-A' represent as shading, normal, and relighting and albedo, respectively.

Figure 8: Normal results with different initializations on the FFHQ dataset [28].

shown in Fig. 8. The coarse exemplar is generated from different pre-training stages, like 1k iterations ('Iter1k'), 50k iterations ('Iter50k'), and 100k iterations ('Iter100k'). We can find that the coarse exemplar has massive artifacts, and after refining by our method, the quality is significantly boosted. By comparing among the exemplars by pre-trianed models with different iterations, our method can always improve the coarse estimation by a large margin. The trained pre-trained model can learn a rough normal, and our model can correct the wrong normal directions and generate high-fidelity normal outputs. The first, second and last lines in Table 3 show normal reconstruction errors with different initializations on the Photoface [56]. The results obtained by different initializing stages of the pre-trained model validate the advance of our proposed method. The pre-trained model learns the distribution of normal roughly, and our method can optimize the feature injection between normal features and structure features to obtain high-fidelity normal.

Architecture. The third, fourth and last rows in Table 3 show the results with different architectures. It is worth mentioning that the condition-based network (***A1***) can achieve the best results in terms of metrics. In other words, they are able to perform well on the testing data of Photoface [56] but poorly on real images, with obvious artifacts, as can be seen in Fig. 9. This is because the distribution of training data is consistent with the test data, while it

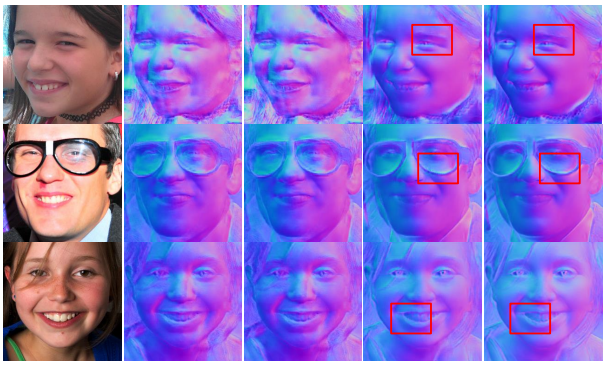

Figure 9: Normal results by different architectures on the FFHQ dataset [28].

Table 3: Comparison in normal reconstruction error with different configurations on the Photoface dataset [56].

Experiments	Mean ± std	$< 20^\circ$	$< 25^\circ$	$< 30^\circ$
Iter1k	16.2±10.1	71.3%	83.0%	90.2%
Iter50k	15.05±9.0	75.1%	86.3%	93.1%
A1	9.6±6.4	94.0%	97.3%	98.6%
A2	11.8±8.1	87.2%	93.5%	96.6%
w/o modulation	11.8±8.3	87.6%	93.1%	96.2%
AdaIN	11.6±8.1	87.9%	93.8%	96.7%
Iter100k/A3/StyleGAN2	11.3±7.7	88.6%	94.4%	97.2%

differs enormously from in-the-wild images. The evaluation metrics of(***A2***) and (***A3***) are not as good as the (***A1***), but their testing results on real data are better than (***A1***). (***A3***) achieves the best results compared to 'Cross-modal' [1].

Feature modulation. Since our implementation involves a feature modulation module [29], we also study the effect of different modulation operations, such as AdaIN [19]. We compare the results used in different operations as given in the 5rd to 7th rows in Table. 3. 'AdaIN' [19], StyleGAN2 [29] and 'w/o modulation' correspond to the trainings with AdaIN feature modulation module, StyleGAN2 feature modulation module and only contact the two features, respectively. Experimental results show that the advantage of our approach used StyleGAN2 [29] feature modulation module is more significant in the coarse-to-fine tasks, and the module allows our method to further improve in the generalization capability.

Applicability of our method. As shown in Fig. 10, we present the results of our method applied to different faces with various skin colors, ages and genders from different datasets. This experiment aims to reveal the generalization ability of our method. The conclusion is consistent with the results shown above. Our method can accurately recover the face normals under different conditions. In addition, our method can optimize from coarse exemplar normals to obtain high-fidelity normals, which indicates that our method can better generalize to unseen data.

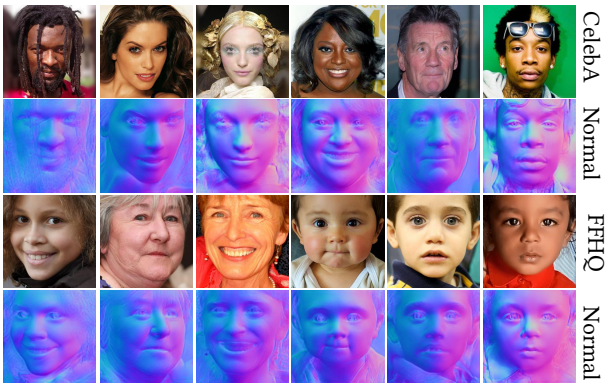

Figure 10: Normal results on various ages, genders, and ethnicities from the CelebA [32] and FFHQ [28] datasets.

Figure 11: Results on low-quality, extreme lighting, and occluded faces.

6 CONCLUDING REMARKS

In this paper, we built a novel framework to solve the problem of high-fidelity face normal estimation. Our method is inspired by the exemplar-based learning and utilizes a coarse exemplar normal as guidance to produce a fine-grained high-quality normal. The framework first converts the coarse exemplar normal into normal features to generate robust results by the feature modulation. This mechanism endows our approach with promising visual quality as well as strong generalization abilities to apply on out-of-distribution face images. Detailed qualitative and quantitative evaluations have shown that our method significantly outperforms other SOTA methods. While our method is robust to many challenging scenarios (*e.g.*, face contained wrinkle and beard), we do observe failure cases as shown in Fig. 11. The very low quality (Fig. 11 (a,b)) and extreme lighting condition/shading (Fig. 11 (c,d)) images leads to inaccurate normal reconstructions. And our method fails on the occlusion face images shown in Fig. 11 (e,f). We notice that these mentioned unrestricted scenarios are challenging not only to our method, but also to (most) existing schemes. It is desired to develop advanced versions based on our work to further cope with these challenging/unrestricted cases.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China under 2019YFC1521200, National Natural Science Foundation of China under Grant 62172295, and Grant 62072327.

REFERENCES

[1] Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. 2020. Cross-modal deep face normals with deactivable skip connections. In *CVPR*. 4979–4989.

[2] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo Masi. 2011. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. 79–80.

[3] Aayush Bansal, Bryan Russell, and Abhinav Gupta. 2016. Marr revisited: 2d-3d alignment via surface normal prediction. In *CVPR*. 5965–5974.

[4] Jonathan T Barron and Jitendra Malik. 2011. High-frequency shape and albedo from shading using natural image statistics. In *CVPR*. IEEE, 2521–2528.

[5] Jonathan T Barron and Jitendra Malik. 2012. Shape, albedo, and illumination from a single image of an unknown object. In *CVPR*. IEEE, 334–341.

[6] Jonathan T Barron and Jitendra Malik. 2014. Shape, illumination, and reflectance from shading. *TPAMI* 37, 8 (2014), 1670–1687.

[7] Chaofeng Chen, Dihong Gong, Hao Wang, Zhifeng Li, and Kwan-Yee K. Wong. 2020. Learning Spatial Attention for Face Super-Resolution. *TIP*.

[8] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. 2015. Deep colorization. In *ICCV*. 415–423.

[9] Nikolai Chinaev, Alexander Chigorin, and Ivan Laptev. 2018. Mobileface: 3D face reconstruction with efficient cnn regression. In *ECCVW*. 0–0.

[10] Cheng Deng, Erkun Yang, Tongliang Liu, Jie Li, Wei Liu, and Dacheng Tao. 2019. Unsupervised semantic-preserving adversarial hashing for image search. *TIP* 28, 8 (2019), 4032–4044.

[11] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. 2017. Learning diverse image colorization. In *CVPR*. 6837–6845.

[12] Berk Dogan, Shuhang Gu, and Radu Timofte. 2019. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*. 0–0.

[13] Ady Ecker and Allan D Jepson. 2010. Polynomial shape from shading. In *CVPR*. IEEE, 145–152.

[14] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*. 534–551.

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *CVPR*. 2414–2423.

[16] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. 2018. Deep exemplar-based colorization. *TOG* 37, 4 (2018), 1–16.

[17] Berthold KP Horn. 1975. Obtaining shape from shading information. *The psychology of computer vision* (1975), 115–155.

[18] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *CVPR*. 5197–5206.

[19] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*. 1501–1510.

[20] Xinyu Huang, Jizhou Gao, Liang Wang, and Ruigang Yang. 2007. Examplar-based shape from shading. In *3DIM*. IEEE, 349–356.

[21] Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*. Springer, 492–518.

[22] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *TOG)* 36, 4 (2017), 1–14.

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*. 1125–1134.

[24] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *ICCV*. 1031–1039.

[25] David W Jacobs and Ronen Basri. 2005. Lambertian reflectance and linear subspaces. US Patent 6,853,745.

[26] Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. 2021. StyleCariGAN: caricature generation via StyleGAN feature map modulation. *TOG* 40, 4 (2021), 1–16.

[27] Taewon Kang. 2021. Multiple GAN Inversion for Exemplar-based Image-to-Image Translation. In *ICCV*. 3515–3522.

[28] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*. 4401–4410.

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *CVPR*. 8110–8119.

[30] Diederik P Kingma and Ba J Adam. 2020. A method for stochastic optimization. arXiv preprint arXiv: 14126980. 2014. *Cited on* (2020), 50.

[31] Iasonas Kokkinos. 2017. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*. 6129–6138.

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. 3730–3738.

[33] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. 2017. Deep photo style transfer. In *CVPR*. 4990–4998.

[34] Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. 2007. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. *Rendering Techniques* 2007, 9 (2007), 10.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS* 32 (2019).

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *CVPR*. 2536–2544.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 234–241.

[38] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*. 397–403.

[39] Matan Sela, Elad Richardson, and Ron Kimmel. 2017. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*. 1576–1585.

[40] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. 2018. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *CVPR*. 6296–6305.

[41] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. 2017. Neural face editing with intrinsic image disentangling. In *CVPR*. 5541–5550.

[42] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[43] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gerard Medioni. 2018. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*. 3935–3944.

[44] Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards high-fidelity nonlinear 3D face morphable model. In *CVPR*. 1126–1135.

[45] Luan Tran and Xiaoming Liu. 2018. Nonlinear 3d face morphable model. In *CVPR*. 7346–7355.

[46] Luan Tran and Xiaoming Liu. 2019. On learning 3d face morphable model from in-the-wild images. *TPAMI* 43, 1 (2019), 157–171.

[47] George Trigeorgis, Patrick Snape, Iasonas Kokkinos, and Stefanos Zafeiriou. 2017. Face normals" in-the-wild" using fully convolutional networks. In *CVPR*. 38–47.

[48] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. 2017. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *CVPR*. 5163–5172.

[49] Ying Xiong, Ayan Chakrabarti, Ronen Basri, Steven J Gortler, David W Jacobs, and Todd Zickler. 2014. From shading to local shape. *TPAMI* 37, 1 (2014), 67–79.

[50] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. 2020. Stylization-based architecture for fast deep exemplar colorization. In *CVPR*. 9363–9372.

[51] Dawei Yang and Jia Deng. 2018. Shape from shading through shape evolution. In *CVPR*. 3781–3790.

[52] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. 2017. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, Vol. 31.

[53] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. 2017. Semantic image inpainting with deep generative models. In *CVPR*. 5485–5493.

[54] Baosheng Yu and Dacheng Tao. 2021. Heatmap Regression via Randomized Rounding. *TPAMI* (2021).

[55] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *CVPR*. 5505–5514.

[56] Stefanos Zafeiriou, Mark Hansen, Gary Atkinson, Vasileios Argyriou, Maria Petrou, Melvyn Smith, and Lyndon Smith. 2011. The photoface database. In *CVPRW*. IEEE, 132–139.

[57] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. 2021. Unbalanced feature transport for exemplar-based image translation. In *CVPR*. 15028–15038.

[58] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. 2019. Deep exemplar-based video colorization. In *CVPR*. 8052–8061.

[59] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. 2020. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*. 5143–5153.

[60] Zhenyu Zhang, Yanhao Ge, Renwang Chen, Ying Tai, Yan Yan, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. 2021. Learning to Aggregate and Personalize 3D Face from In-the-Wild Photo Collection. In *CVPR*. 14214–14224.

[61] Hengyuan Zhao, Wenhao Wu, Yihao Liu, and Dongliang He. 2021. Color2Embed: Fast Exemplar-Based Image Colorization using Color Embeddings. *arXiv preprint arXiv:2106.08017* (2021).

[62] Haitian Zheng, Minghao Guo, Haoqian Wang, Yebin Liu, and Lu Fang. 2017. Combining exemplar-based approach and learning-based approach for light field super-resolution using a hybrid imaging system. In *ICCVW*. 2481–2486.

[63] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. 2016. Face alignment across large poses: A 3d solution. In *CVPR*. 146–155.